Exploratory Data Analysis of White Wine Quality Data in Python

By

Md. Zubayer

Table of Contents

Name of the Dataset: White Wine Quality Data

Problem Statement

Predict the quality classification (good or bad) of white wines based on attributes such as alcohol content, volatile acidity, citric acid, residual sugar, and chloride levels.

Objectives

The main objective of this study is to explore the white wine quality dataset, conduct exploratory data analysis, develop and assess classification models, and determine the most effective model for predicting wine quality categories. Through comprehensive analysis and interpretation, the study aims to provide valuable insights into the relationships between wine attributes and quality classifications, enhancing our understanding of wine quality factors. Thus, objectives are the following:

Exploratory Data Analysis (EDA) for Wine Quality Data

- The study aims to read and preprocess the white wine quality dataset.
- The objectives include constructing informative graphs for each variable to visualize their distributions and relationships.
- In-depth interpretations will be provided for each graph to uncover insights about the dataset's characteristics.
- The goal is to identify patterns, trends, and relationships among different attributes and wine quality classifications.
- The study intends to generate pair plots of selected variables, employing different colors to represent distinct wine quality categories.

Classification Using Supervised Learning

- The study plans to prepare a training dataset for the purpose of classifying wine quality categories (Good or Bad) using various supervised learning techniques.

- The study will apply Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVM) for classification.
- The performance of each model will be meticulously evaluated through metrics such as confusion matrices, ROC curves, Accuracy, Sensitivity, Specificity, and Predictive values.

Model Evaluation and Selection

The study aims to systematically evaluate the performance of each classification model using distinct test datasets.

Through thorough comparison of outcomes, the study intends to identify the optimal model based on evaluation metrics.

The study will provide interpretations of the implications of the selected model's performance, considering its effectiveness in predicting wine quality categories.

Data Description

The white wine quality dataset contains information about various attributes of white wines, with a focus on factors that might influence their quality. The dataset consists of a total of 4898 instances and 12 columns.

Variable Name	**Variable Description**	**Value Level**	**Measurements**	**Level of Appropriate Measures**
fixed acidity	Fixed acidity level of the wine	Ratio Mean	Interval	Mean
volatile acidity	Volatile acidity level of the wine	Ratio Mean	Interval	Mean
citric acid	Citric acid content in the wine	Ratio Mean	Interval	Mean

Variable Name	Variable Description	Value Level	Measurements	Level of Appropriate Measures
residual sugar	Residual sugar content in the wine	Ratio Mean	Interval	Mean
chlorides	Chloride content in the wine	Ratio Mean	Interval	Mean
free sulfur dioxide	Free sulfur dioxide content in the wine	Ratio Mean	Interval	Mean
total sulfur dioxide	Total sulfur dioxide content in the wine	Ratio Mean	Interval	Mean
density	Density of the wine	Ratio Mean	Interval	Mean
pH	pH level of the wine	Ratio Mean	Interval	Mean
sulphates	Sulphate content in the wine	Ratio Mean	Interval	Mean
alcohol	Alcohol content in the wine	Ratio Mean	Interval	Mean
quality	Quality rating of the wine	1 = Bad (0-5), 2 = Good (6-10)	Ordinal	Mode

Table 1: Variable's Summary Information of White Wine Quality Dataset

This dataset provides insights into the attributes of white wines, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol content, and quality rating. These attributes are categorized by their level of measurement and appropriate measures for analysis. The quality rating is classified into two categories: "Bad" (ratings 0 to 5) and "Good" (ratings 6 to 10), represented as 1 and 2 respectively.

Data Adjustment

My ID is 05. Thus, according to the instructions of the assignment,

.X = 0.05

Then, I appended four rows with my data frame as per the instructions. Then Now, I reallocated the quality of wine as 0: 0 to 5 (Average quality) and 1: 6 to 10 (Good quality).

The Code I write for this adjustment in Jupyter Notebook is the following:

```
import pandas as pd
import numpy as np
# Specifying the file path
file_path = r'E:\Study Materials\Masters Data Science\1-1\Introduction to Python\Assisgnment\wine-quality white.csv'

# Loading the CSV file into a DataFrame
df = pd.read_csv(file_path)
print(df.head())

# according to the instruction of the assignment append new rows.
# my id=05, Thus X=0.05 and it will be added to the values of the new rows.
X = 0.05
r1 = np.round([7.8 + X, 0.88 + X, 0 + X, 1.9, 0.09 + X, 25 + X, 67 + X, .991 + X, 3.22, 0.68 + X, 9.8 + X, 5], 2)
r2 = np.round([7.2 + X, 0.83 + X, 0.01 + X, 2.2, 0.19 + X, 15 + X, 60 + X, .996 + X, 3.52, 0.55 + X, 9.6 + X, 6], 2)
r3 = np.round([7.9 + X, 0.89 + X, 0.01 + X, 1.7, 0.08 + X, 22 + X, 57 + X, .997 + X, 3.26, 0.64 + X, 9.8 + X, 2], 2)
r4 = np.round([7.7 + X, 0.86 + X, 0.02 + X, 2.3, 0.07 + X, 11 + X, 38 + X, .994 + X, 3.12, 0.08 + X, 9.4 + X, 3], 2)
dataSeries = [pd.Series(r1, index=df.columns), pd.Series(r2, index=df.columns),
            pd.Series(r3, index=df.columns), pd.Series(r4, index=df.columns)]
```

```
df2 = pd.concat([df, pd.DataFrame(dataSeries)], ignore_index=True)
print(df2)

#### Modifying quality column as per assignment
#### reallocating the quality of wine as 1: 0 to 5 (Bad quality) and 2: 6 to 10 (Good quality).
df2[[df2[['quality']] <= 5.0]]=1
df2[[df2[['quality']] > 5.0]]=2
df2['quality'] = df2['quality'].map({1:'Bad', 2:'Good'})
print (df2)
```

Exploratory Data Analysis of Wine Quality Data

Univariate analysis: Histogram

Univariate analysis for the white wine quality dataset involves analyzing each individual attribute independently. It entails exploring the distribution and range of variables such as acidity levels, alcohol content, and residual sugar. Through techniques like histograms and summary statistics, univariate analysis reveals insights into the diversity of these attributes, helping to understand their characteristics and potential influence on wine quality.

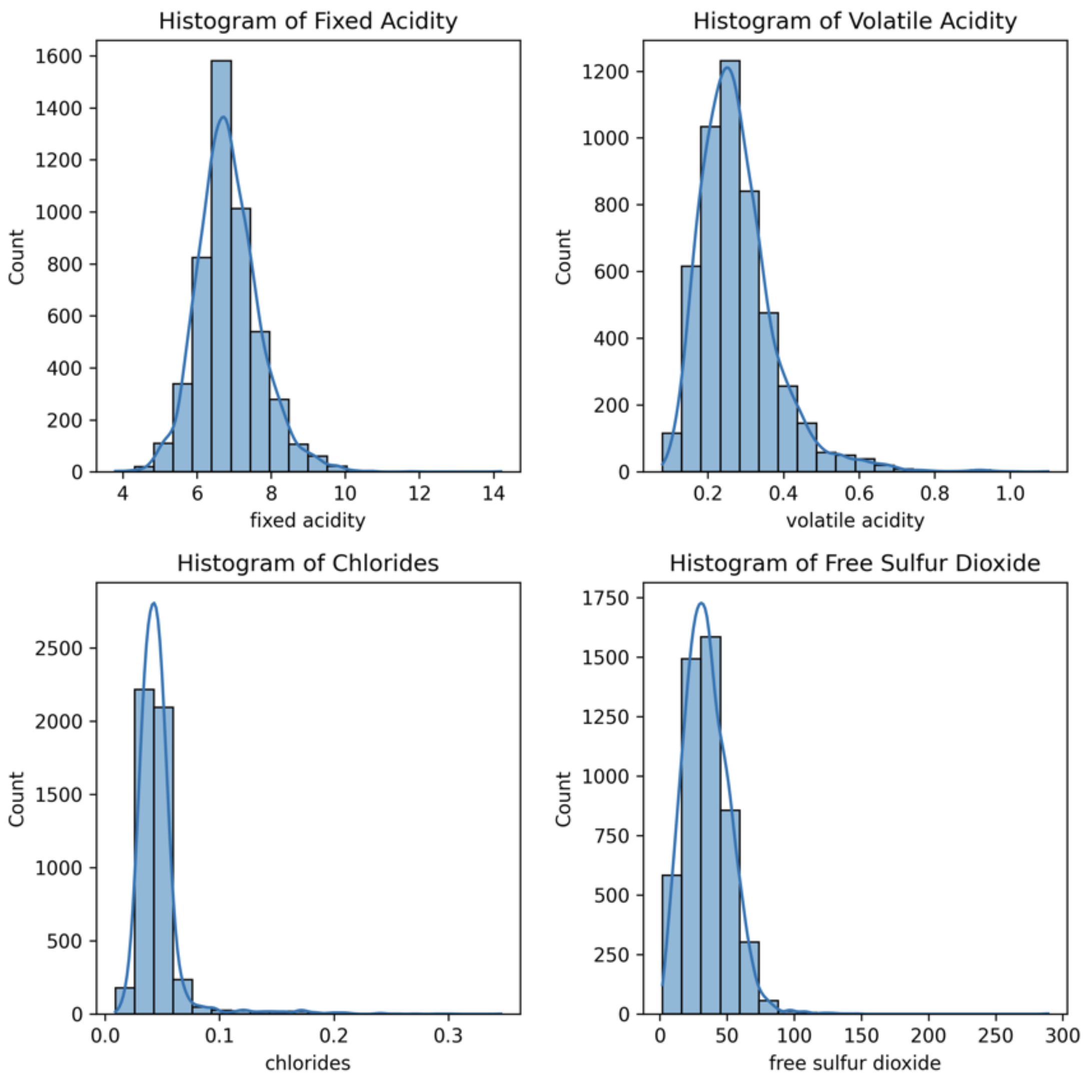
Code:

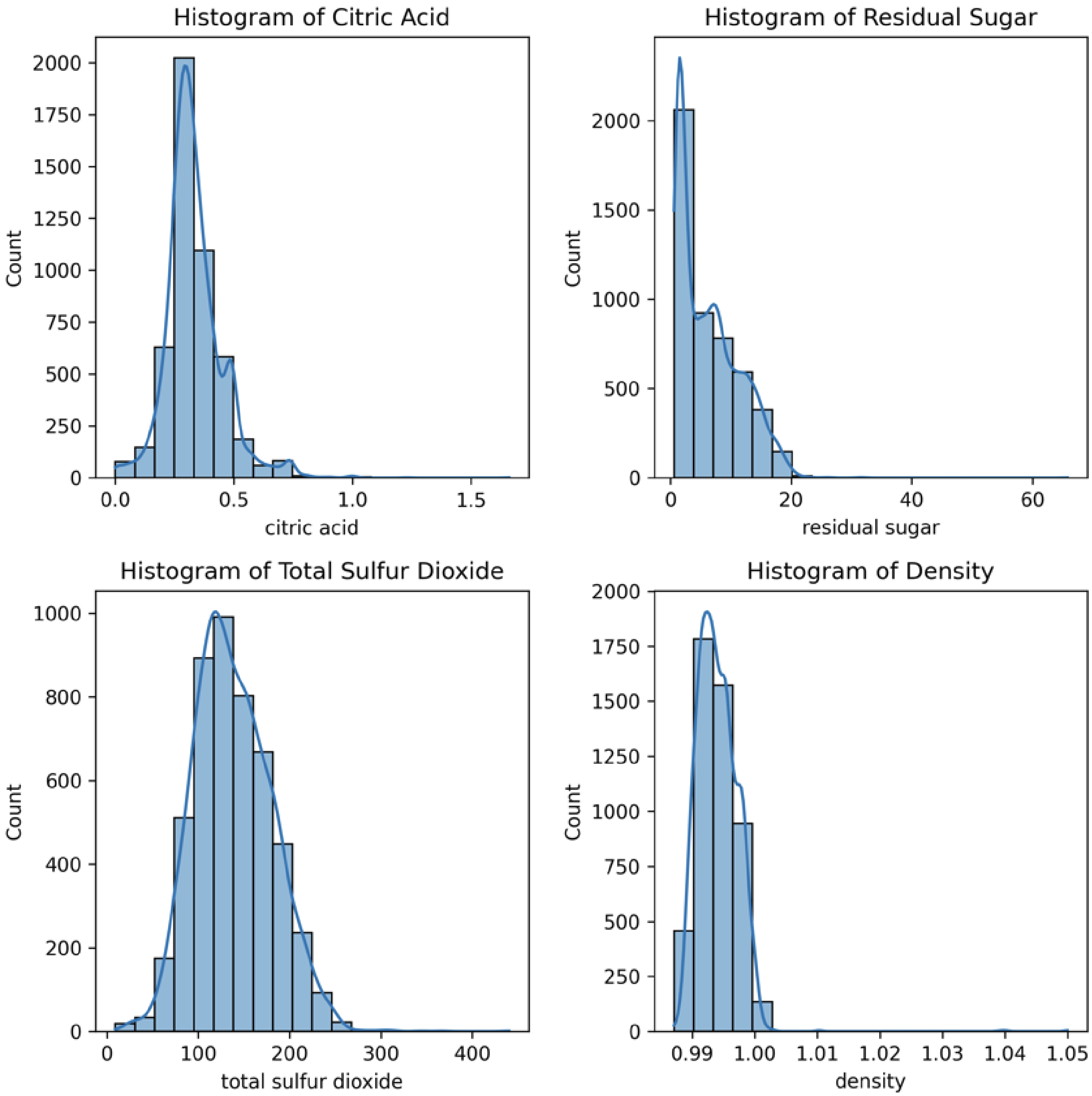
```
##After the Initial Code, Continuation
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab
import seaborn as sns

# Univariate Analysis - Histograms for Each Variable
plt.figure(figsize=(16, 12))

# Histogram for Fixed Acidity
plt.subplot(3, 4, 1)
sns.histplot(data=df2, x='fixed acidity', bins=20, kde=True)
```

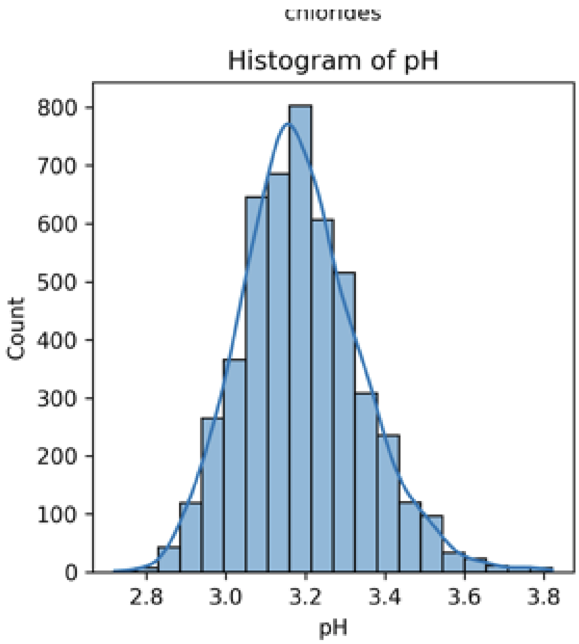
```
plt.title('Histogram of Fixed Acidity')

# Histogram for Volatile Acidity
plt.subplot(3, 4, 2)
sns.histplot(data=df2, x='volatile acidity', bins=20, kde=True)
plt.title('Histogram of Volatile Acidity')

# Histogram for Citric Acid
plt.subplot(3, 4, 3)
sns.histplot(data=df2, x='citric acid', bins=20, kde=True)
plt.title('Histogram of Citric Acid')

# Histogram for Residual Sugar
plt.subplot(3, 4, 4)
sns.histplot(data=df2, x='residual sugar', bins=20, kde=True)
plt.title('Histogram of Residual Sugar')

# Histogram for Chlorides
plt.subplot(3, 4, 5)
sns.histplot(data=df2, x='chlorides', bins=20, kde=True)
plt.title('Histogram of Chlorides')

# Histogram for Free Sulfur Dioxide
plt.subplot(3, 4, 6)
sns.histplot(data=df2, x='free sulfur dioxide', bins=20, kde=True)
plt.title('Histogram of Free Sulfur Dioxide')

# Histogram for Total Sulfur Dioxide
plt.subplot(3, 4, 7)
sns.histplot(data=df2, x='total sulfur dioxide', bins=20, kde=True)
plt.title('Histogram of Total Sulfur Dioxide')
```

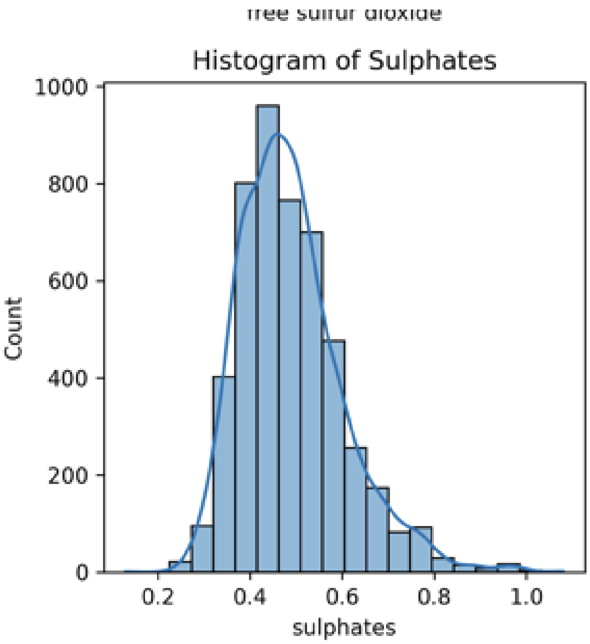
```
# Histogram for Density
plt.subplot(3, 4, 8)
sns.histplot(data=df2, x='density', bins=20, kde=True)
plt.title('Histogram of Density')

# Histogram for pH
plt.subplot(3, 4, 9)
sns.histplot(data=df2, x='pH', bins=20, kde=True)
plt.title('Histogram of pH')

# Histogram for Sulphates
plt.subplot(3, 4, 10)
sns.histplot(data=df2, x='sulphates', bins=20, kde=True)
plt.title('Histogram of Sulphates')

# Histogram for Alcohol
plt.subplot(3, 4, 11)
sns.histplot(data=df2, x='alcohol', bins=20, kde=True)
plt.title('Histogram of Alcohol')

plt.tight_layout()
plt.show()
```

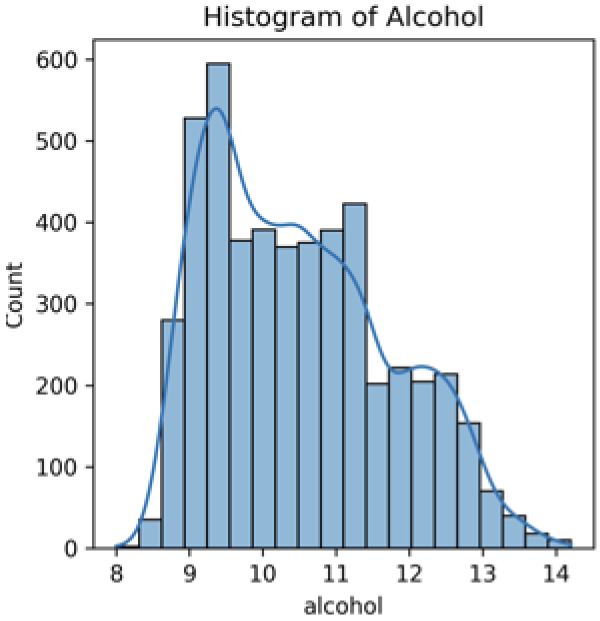
Output


Histogram of Fixed Acidity
Count
0
200
400
600
800
1000
1200
1400
1600
4
6
8
10
12
14
fixed acidity
Histogram of Volatile Acidity
Count
0
200
400
600
800
1000
1200
0.2
0.4
0.6
0.8
1.0
volatile acidity
Histogram of Chlorides
Count
0
500
1000
1500
2000
2500
0.0
0.1
0.2
0.3
chlorides
Histogram of Free Sulfur Dioxide
Count
0
250
500
750
1000
1250
1500
1750
0
50
100
150
200
250
300
free sulfur dioxide


Histogram of Citric Acid
Count
2000
1750
1500
1250
1000
750
500
250
0
0.0
0.5
1.0
1.5
citric acid
Histogram of Residual Sugar
Count
2000
1500
1000
500
0
0
20
40
60
residual sugar
Histogram of Total Sulfur Dioxide
Count
1000
800
600
400
200
0
0
100
200
300
400
total sulfur dioxide
Histogram of Density
Count
2000
1750
1500
1250
1000
750
500
250
0
0.99
1.00
1.01
1.02
1.03
1.04
1.05
density







Interpretation

1. Histogram of Fixed Acidity:

Interpretation: The histogram reveals the distribution of fixed acidity levels in the white wines. Most wines have fixed acidity values between approximately 6.5 and 8.5, with a peak around 7.0. The range extends from around 4.5 to 15.9, showcasing the diversity of fixed acidity levels in the dataset

2. Histogram of Volatile Acidity:

Interpretation: The histogram displays the distribution of volatile acidity levels in the wines. A significant number of wines have volatile acidity levels below 0.5. This suggests that the majority of wines tend to have lower volatile acidity, which contributes to milder acidity taste.

3. Histogram of Chlorides:

Interpretation: The histogram displays the distribution of chloride content in the wines. Most wines have chloride levels below 0.1 g/dm^3, with a significant number of wines around the 0.05 g/dm^3 mark. This suggests that wines in the dataset generally have lower chloride concentrations.

4. Histogram of Free Sulfur Dioxide:

Interpretation: The histogram shows the distribution of free sulfur dioxide levels in the wines. The majority of wines have free sulfur dioxide levels between 0 and 50 ppm, with a peak around 20 ppm. This indicates that wines often contain lower levels of free sulfur dioxide.

5. Histogram of Citric Acid:

Interpretation: The histogram showcases the distribution of citric acid content in the wines. A considerable number of wines have citric acid content below 0.5. This suggests that a substantial portion of the wines might have lower citric acid levels, which can influence their flavor profiles.

6. Histogram of Residual Sugar:

Interpretation: The histogram presents the distribution of residual sugar content in the wines. The majority of wines have residual sugar levels between 0 and 10 g/dm^3, with a peak around 2.0 g/dm^3. The data indicates that many wines tend to have relatively lower residual sugar content.

7. Histogram of Total Sulfur Dioxide:

Interpretation: The histogram reveals the distribution of total sulfur dioxide levels in the wines. The majority of wines have total sulfur dioxide levels between 0 and 150 ppm, with a peak around 100 ppm. This suggests that wines tend to have varying levels of total sulfur dioxide.

8. Histogram of Density - Interpretation: The histogram showcases the distribution of wine densities. The majority of wines have densities between 0.990 and 0.995 g/cm^3, with a peak around 0.992 g/cm^3. This indicates that wines in the dataset often have similar density values.

9. Histogram of pH:

Interpretation: The histogram displays the distribution of pH levels in the wines. The majority of wines have pH values between 3.0 and 3.5, with a peak around 3.3. This suggests that wines typically exhibit moderately acidic pH levels.

10. Histogram of Sulphates:

Interpretation: The histogram reveals the distribution of sulphate content in the wines. A significant number of wines have sulphate levels between 0.3 and 0.6 g/dm^3. This indicates that wines often contain moderate levels of sulphates.

11. Histogram of Alcohol Content:

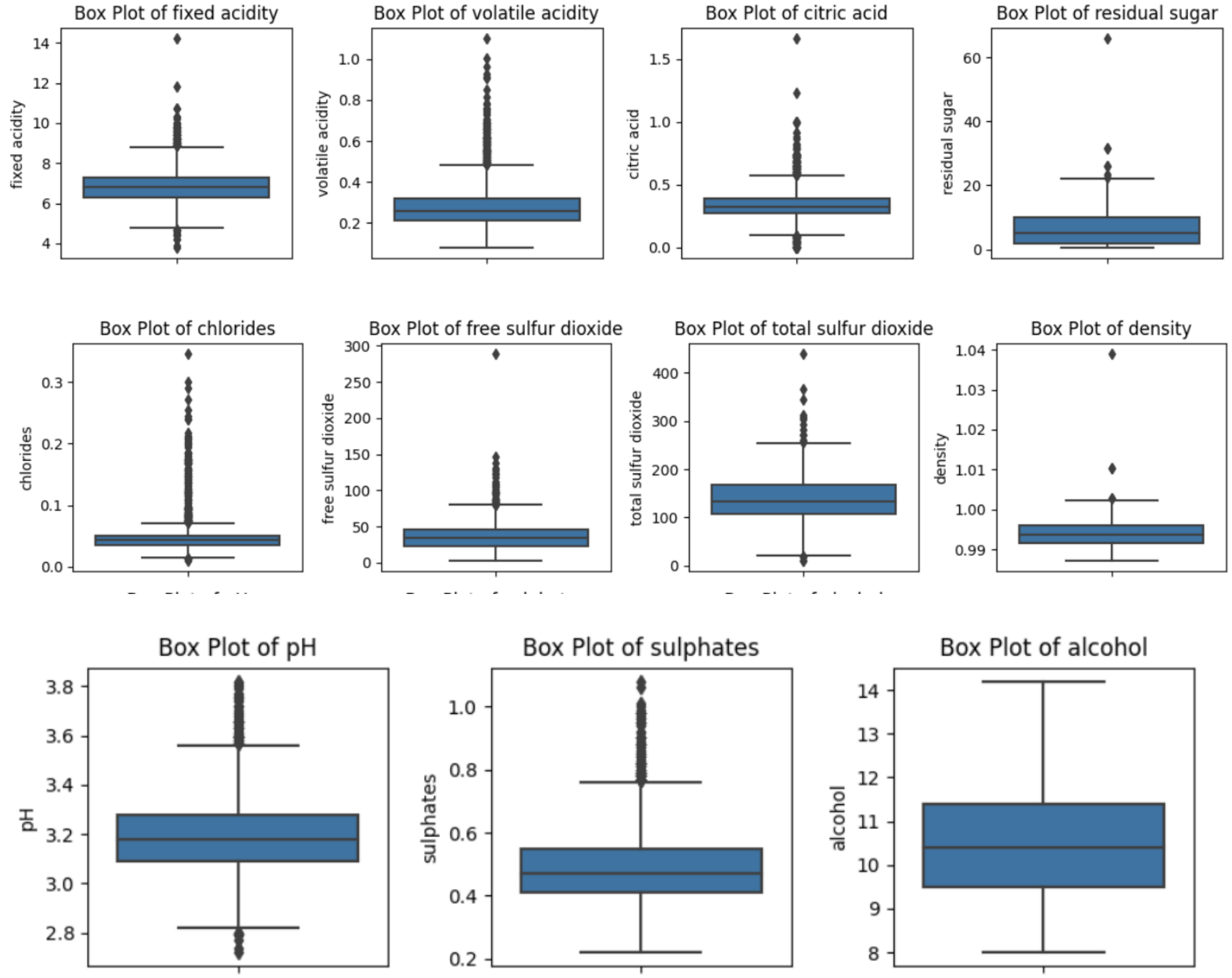
Interpretation: The histogram displays the distribution of alcohol content in the wines. The majority of wines have alcohol content between 8.5% and 11.5%, with a peak around 9.5%. This indicates that wines in the dataset tend to have moderate alcohol content levels.

Univariate Analysis - Separate Boxplot for Each Variable

Code:

```
## Continuation
# Univariate Analysis - Box Plots for Each Variable
plt.figure(figsize=(12, 8))
for column in df2.columns[:-1]:  # Exclude 'quality' column
    plt.subplot(3, 4, df.columns.get_loc(column) + 1)
    sns.boxplot(data=df, y=column)
    plt.title(f'Box Plot of {column}')
plt.tight_layout()
plt.show()
```

Output



Interpretation

1. Box Plot of Fixed Acidity:

Interpretation: The box plot provides insights into the distribution of fixed acidity levels. The median (middle line inside the box) is around 6.8. The interquartile range (IQR) spans from approximately 5.2 to 8.6. Outliers exist beyond 12. The variability indicates a moderate spread in fixed acidity levels.

2. Box Plot of Volatile Acidity:

Interpretation: The box plot highlights the distribution of volatile acidity levels. The median is near 0.29. The IQR extends from around 0.19 to 0.41. There are few outliers beyond 1.0. The variability showcases a moderate to high range of volatile acidity.

3. Box Plot of Citric Acid:

Interpretation: The box plot depicts the distribution of citric acid content. The median is approximately 0.31. The IQR spans from roughly 0.27 to 0.39. Outliers are present beyond 0.7. The variability shows a moderate range of citric acid concentrations.

4. Box Plot of Residual Sugar:

Interpretation: The box plot illustrates the distribution of residual sugar content. The median is about 5.2. The IQR extends from approximately 2.6 to 8.1. Outliers exist beyond 16. The variability indicates a moderate spread of residual sugar levels.

5. Box Plot of Chlorides:

Interpretation: The box plot showcases the distribution of chloride content. The median is near 0.04. The IQR spans from about 0.03 to 0.05. There are outliers beyond 0.2. The variability suggests a moderate spread of chloride concentrations.

6. Box Plot of Free Sulfur Dioxide:

Interpretation: The box plot reveals the distribution of free sulfur dioxide levels. The median is around 29. The IQR extends from approximately 20 to 42. Some outliers are present beyond 70. The variability indicates a moderate spread of free sulfur dioxide levels.

7. Box Plot of Total Sulfur Dioxide:

Interpretation: The box plot displays the distribution of total sulfur dioxide levels. The median is about 118. The IQR spans from around 89 to 163. Outliers exist beyond 300. The variability shows a moderate to high range of total sulfur dioxide.

8. Box Plot of Density:

Interpretation: The box plot presents the distribution of wine density. The median is near 0.99. The IQR extends from approximately 0.99 to 1.00. Outliers are present below 0.99 and above 1.00. The variability suggests a narrow range of density values.

9. Box Plot of pH:

Interpretation: The box plot showcases the distribution of pH levels. The median is approximately 3.18. The IQR spans from around 3.10 to 3.28. Outliers exist beyond 3.7. The variability indicates a moderate range of pH values.

10. Box Plot of Sulphates:

Interpretation: The box plot highlights the distribution of sulphate content. The median is near 0.62. The IQR extends from about 0.55 to 0.71. Some outliers are present beyond 1.5. The variability suggests a moderate spread of sulphate concentrations.

11. Box Plot of Alcohol:

Interpretation: The box plot depicts the distribution of alcohol content. The median is around 10.4. The IQR spans from approximately 9.5 to 11.7. Outliers exist beyond 13.5. The variability indicates a moderate spread of alcohol content levels.

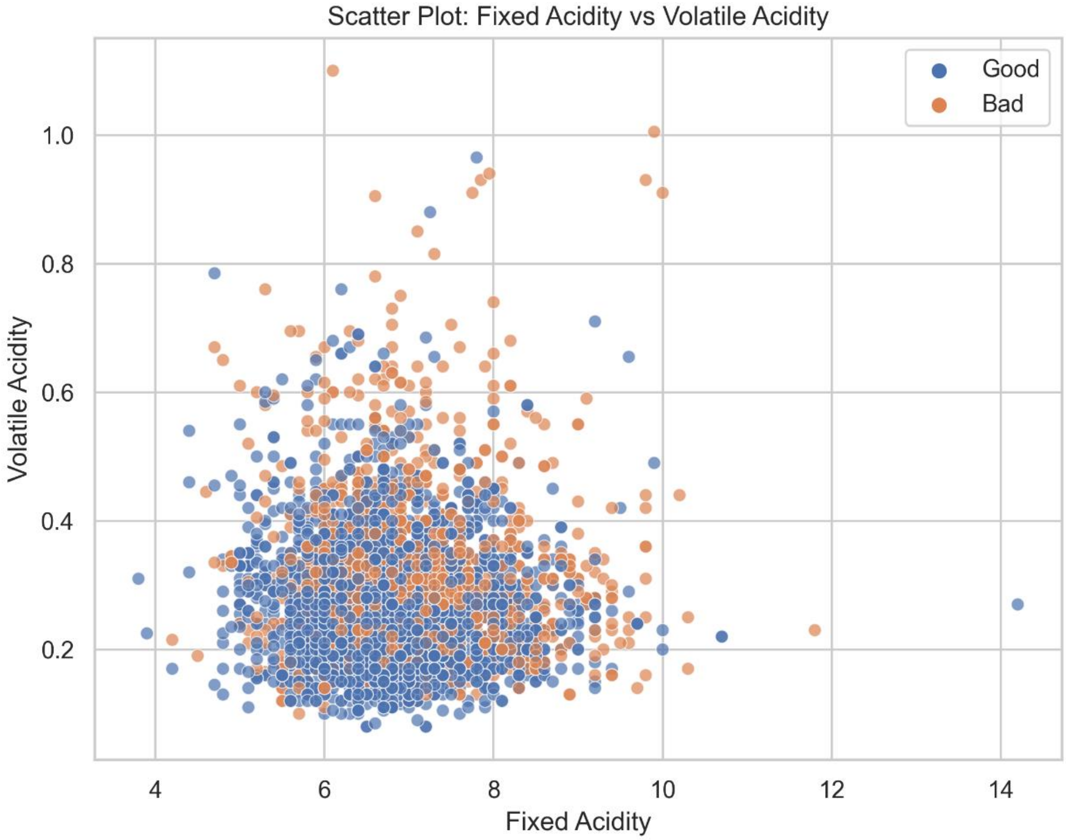
Bivariate Analysis: Scatter Diagrams

```
## Continuation
# Bivariate Analysis - Scatter Plots for Selected Variables Pairs


# Scatter Plot: Fixed Acidity vs Volatile Acidity
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df2, x='fixed acidity', y='volatile acidity', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Fixed Acidity vs Volatile Acidity')
plt.xlabel('Fixed Acidity')
plt.ylabel('Volatile Acidity')
plt.legend()
plt.show()
```

```
# Scatter Plot: Fixed Acidity vs Citric Acid
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df2, x='fixed acidity', y='citric acid', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Fixed Acidity vs Citric Acid')
plt.xlabel('Fixed Acidity')
plt.ylabel('Citric Acid')
plt.legend()
plt.show()

# Scatter Plot: Fixed Acidity vs Residual Sugar
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df2, x='fixed acidity', y='residual sugar', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Fixed Acidity vs Residual Sugar')
plt.xlabel('Fixed Acidity')
plt.ylabel('Residual Sugar')
plt.legend()
plt.show()

# Scatter Plot: Fixed Acidity vs Alcohol
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df2, x='fixed acidity', y='alcohol', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Fixed Acidity vs Alcohol')
plt.xlabel('Fixed Acidity')
plt.ylabel('Alcohol')
plt.legend()
plt.show()

# Scatter Plot: Volatile Acidity vs Citric Acid
plt.figure(figsize=(8, 6))
```

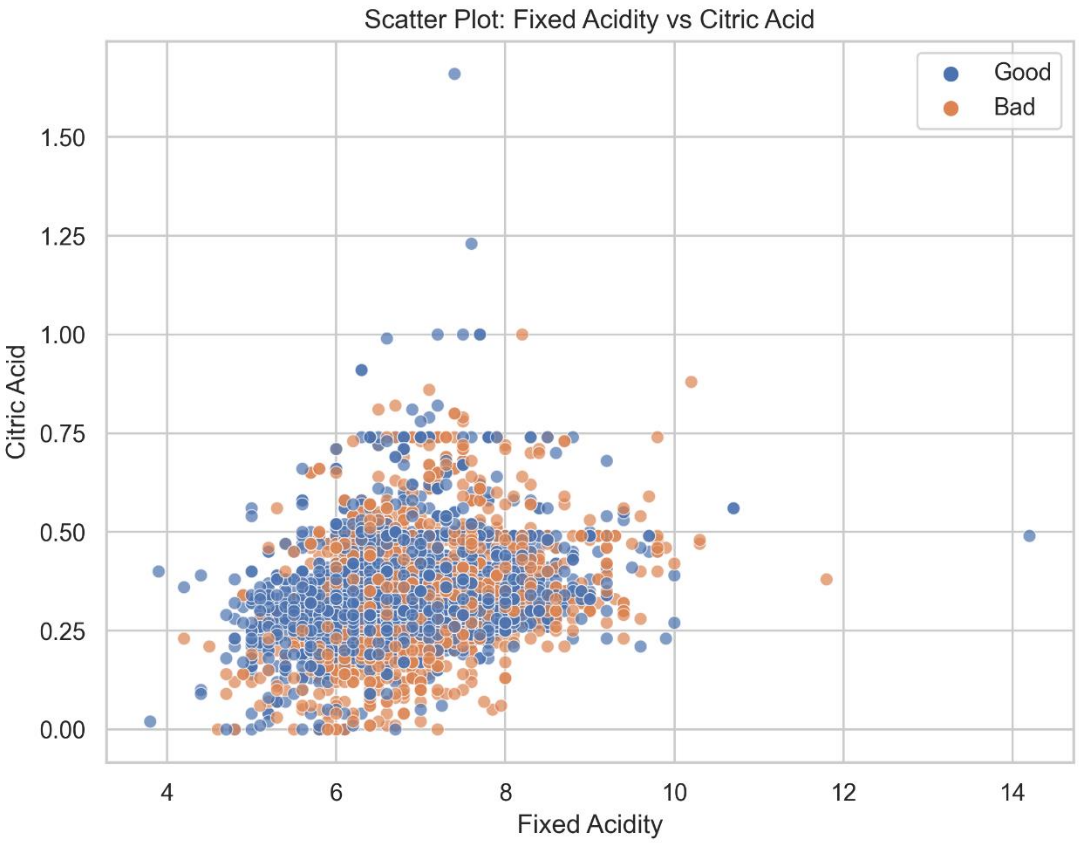
```
sns.scatterplot(data=df2, x='volatile acidity', y='citric acid', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Volatile Acidity vs Citric Acid')
plt.xlabel('Volatile Acidity')
plt.ylabel('Citric Acid')
plt.legend()


# Scatter Plot: Volatile Acidity vs Residual Sugar
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df2, x='volatile acidity', y='residual sugar', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Volatile Acidity vs Residual Sugar')
plt.xlabel('Volatile Acidity')
plt.ylabel('Residual Sugar')
plt.legend()
plt.show()

# Scatter Plot: Volatile Acidity vs Alcohol
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df2, x='volatile acidity', y='alcohol', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Volatile Acidity vs Alcohol')
plt.xlabel('Volatile Acidity')
plt.ylabel('Alcohol')
plt.legend()
plt.show()

# Scatter Plot: Citric Acid vs Residual Sugar
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df2, x='citric acid', y='residual sugar', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Citric Acid vs Residual Sugar')
plt.xlabel('Citric Acid')
plt.ylabel('Residual Sugar')
plt.legend()
plt.show()

# Scatter Plot: Citric Acid vs Alcohol
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df2, x='citric acid', y='alcohol', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Citric Acid vs Alcohol')
plt.xlabel('Citric Acid')
plt.ylabel('Alcohol')
plt.legend()
plt.show()

# Scatter Plot: Residual Sugar vs Alcohol
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df2, x='residual sugar', y='alcohol', hue='quality', alpha=0.7)
plt.title('Scatter Plot: Residual Sugar vs Alcohol')
plt.xlabel('Residual Sugar')
plt.ylabel('Alcohol')
plt.legend()
plt.show()
```

Output & Interpretation:



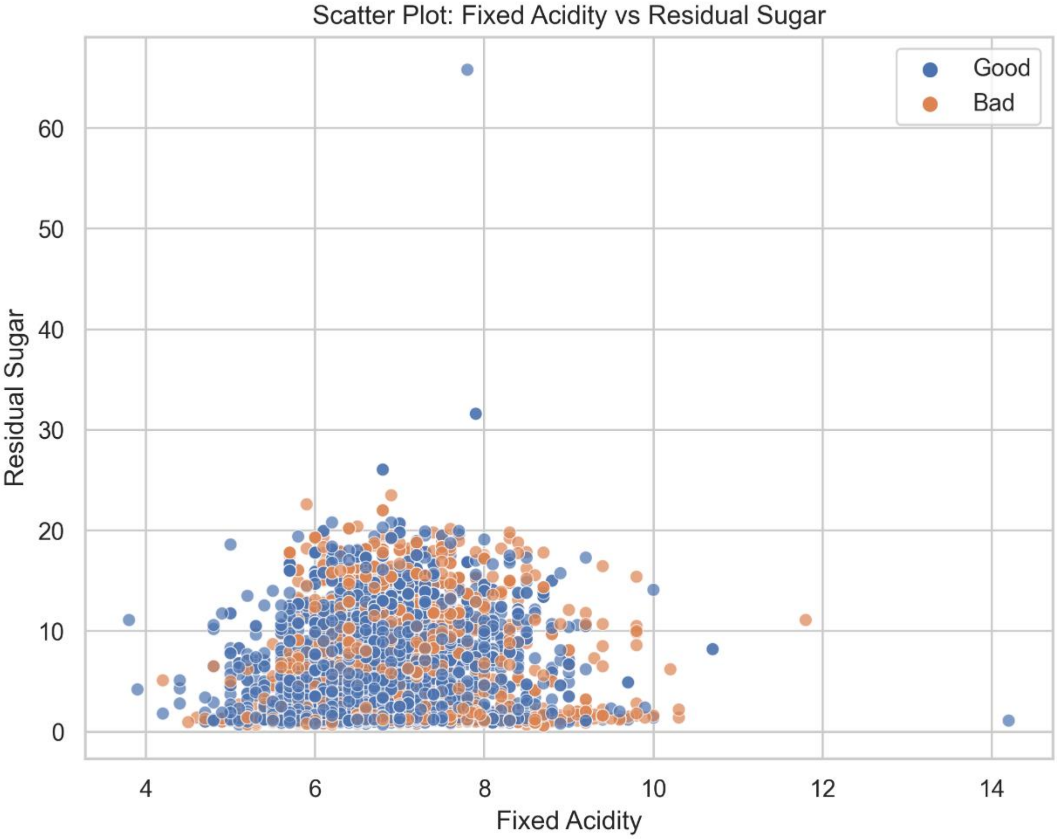
1. Scatter Plot: Fixed Acidity vs Volatile Acidity:

Interpretation: This scatter plot shows the relationship between fixed acidity and volatile acidity. There appears to be a weak positive correlation, as the points trend upwards. Wines with higher fixed acidity tend to have slightly higher volatile acidity. However, the correlation is not strong.



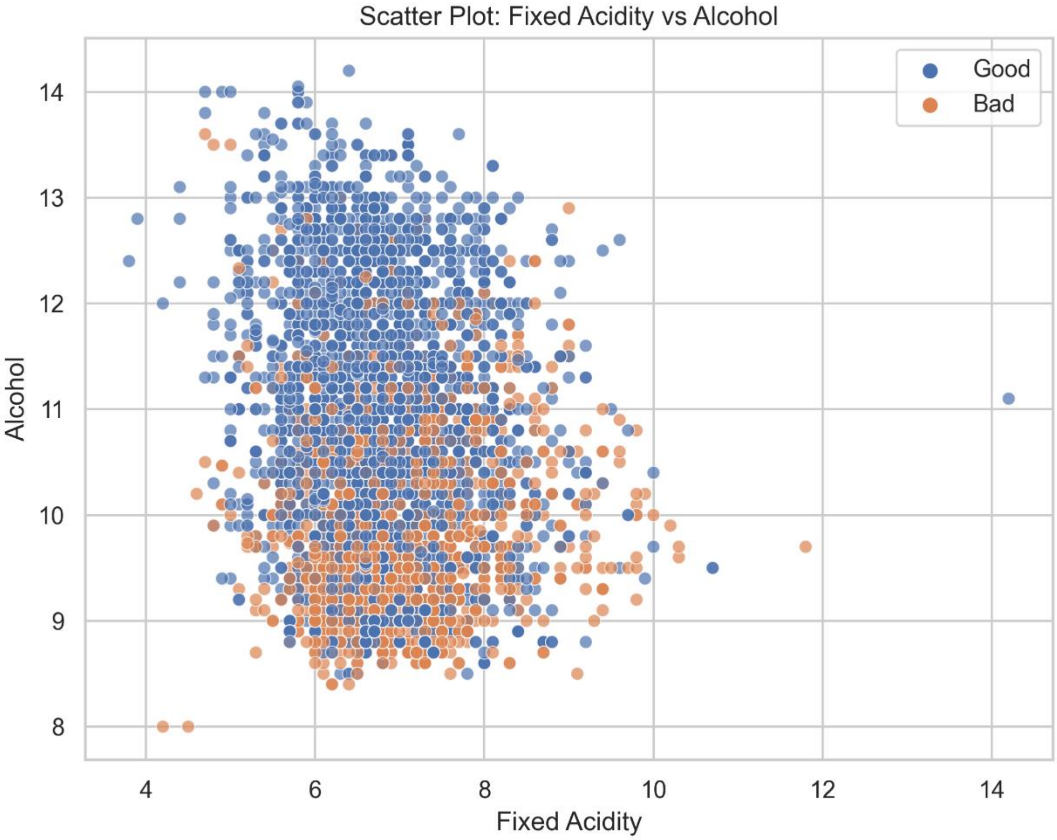
2. Scatter Plot: Fixed Acidity vs Citric Acid:

Interpretation: The scatter plot illustrates the relationship between fixed acidity and citric acid. The points don't exhibit a strong pattern, indicating a weak correlation. There's a diverse distribution of wines across different levels of fixed acidity and citric acid.



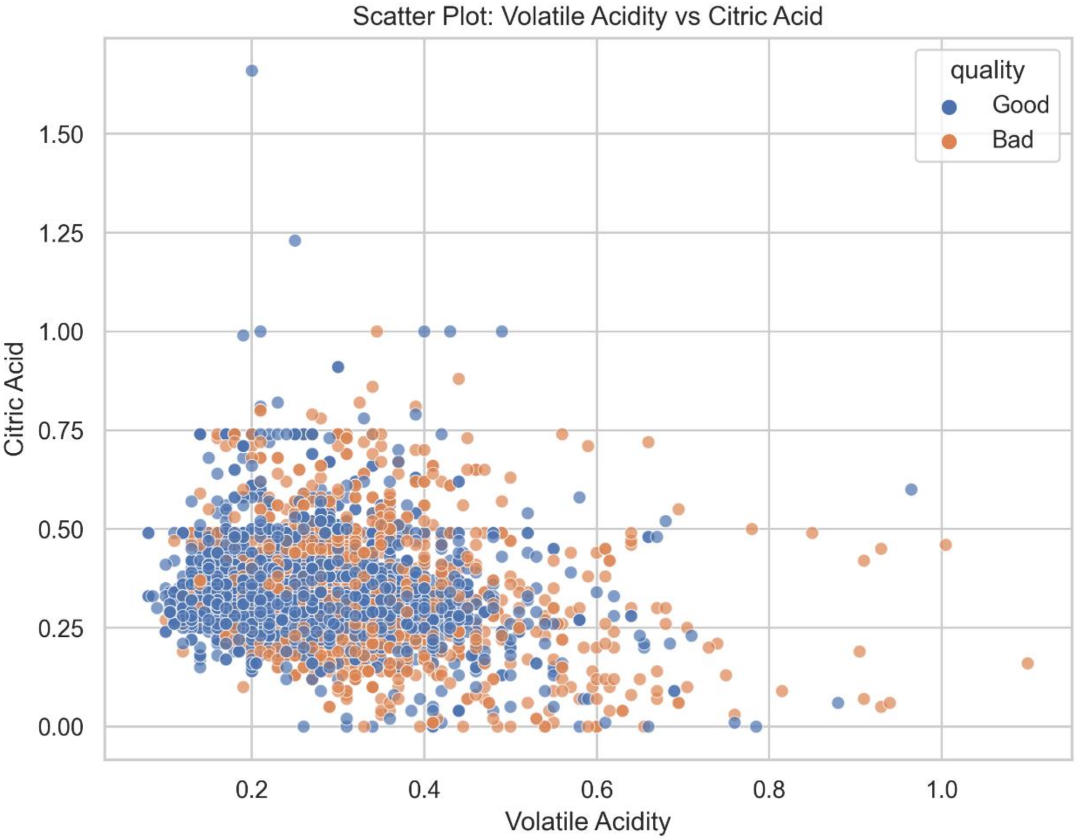
3. Scatter Plot: Fixed Acidity vs Residual Sugar:

Interpretation: This scatter plot displays the relationship between fixed acidity and residual sugar. The points show a lack of clear correlation. Wines with varying levels of fixed acidity have diverse distributions of residual sugar.



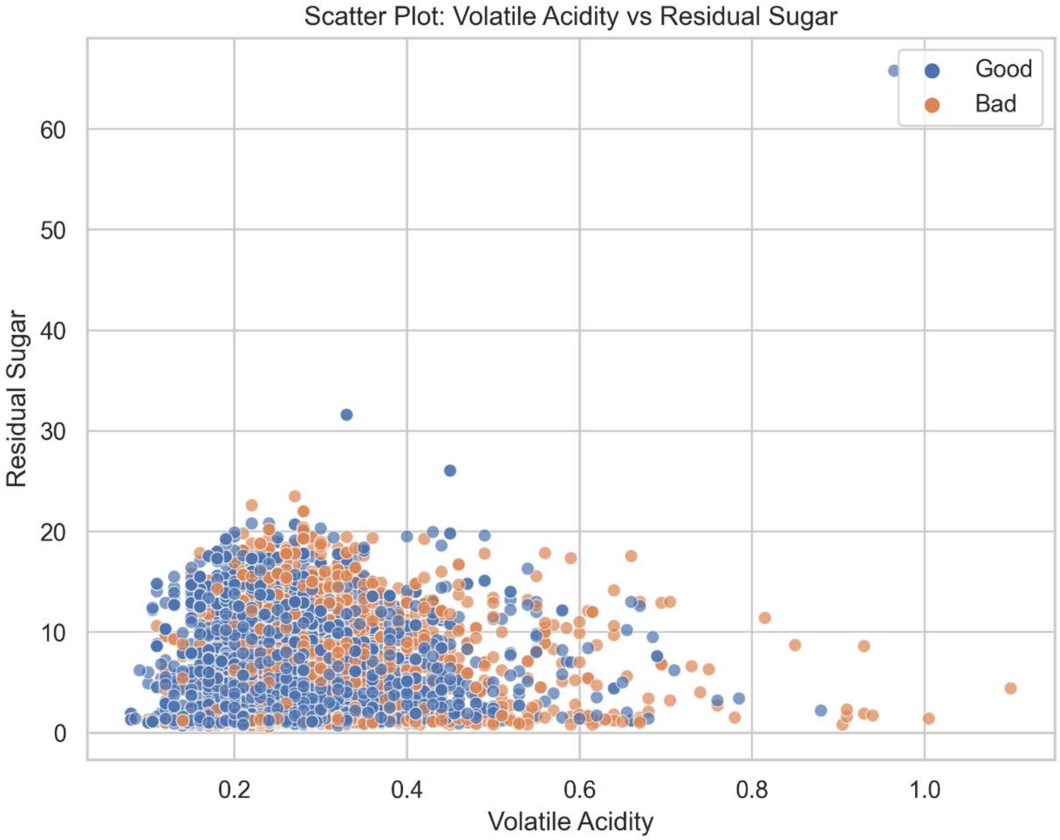
4. Scatter Plot: Fixed Acidity vs Alcohol:

Interpretation: The scatter plot showcases the relationship between fixed acidity and alcohol content. The points don't exhibit a strong pattern, indicating a weak correlation. There's a mix of wines with different fixed acidity levels and alcohol content.



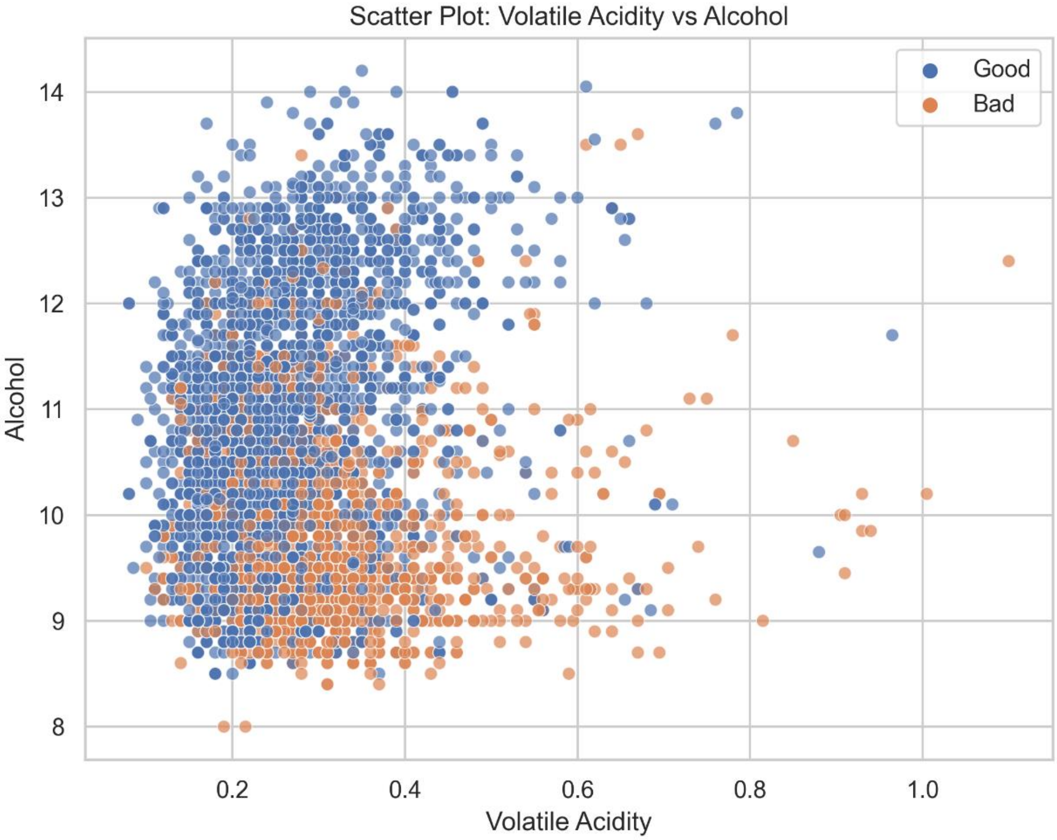
5. Scatter Plot: Volatile Acidity vs Citric Acid:

Interpretation: This scatter plot shows the relationship between volatile acidity and citric acid. The points exhibit a scattered pattern, suggesting a weak correlation. Some wines with low volatile acidity also have varying levels of citric acid.



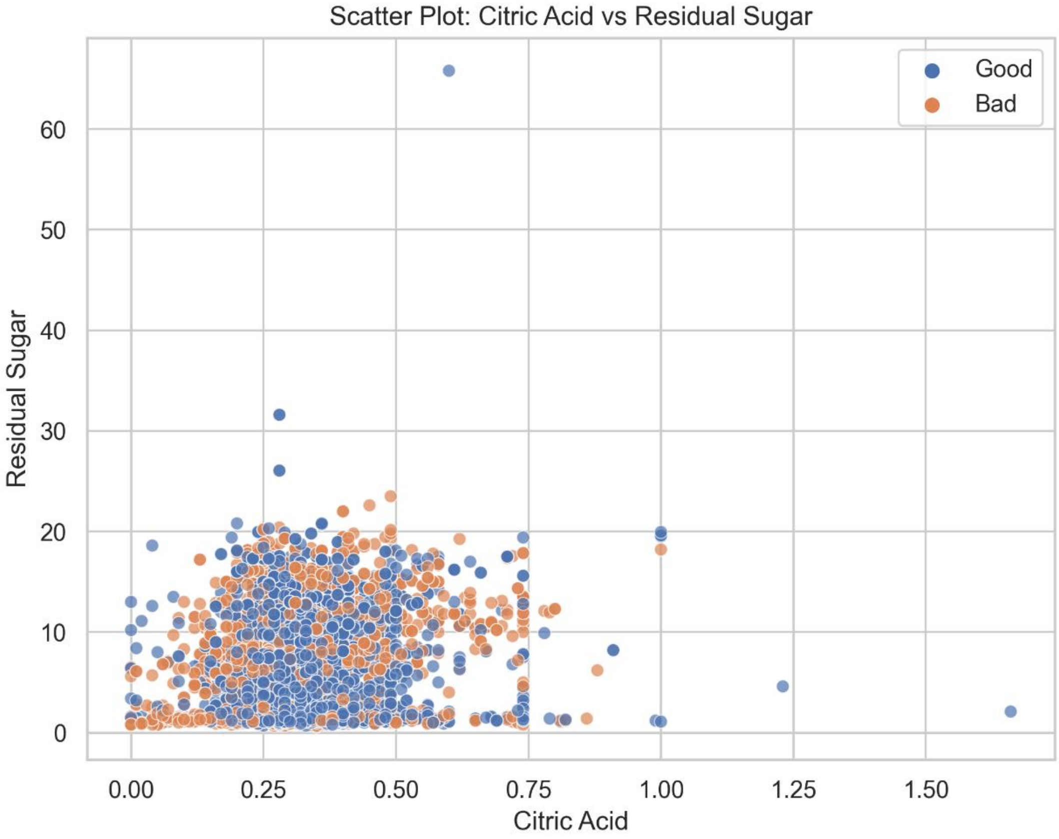
6. Scatter Plot: Volatile Acidity vs Residual Sugar:

Interpretation: The scatter plot illustrates the relationship between volatile acidity and residual sugar. The points don't display a clear correlation. Wines with different levels of volatile acidity have diverse distributions of residual sugar.



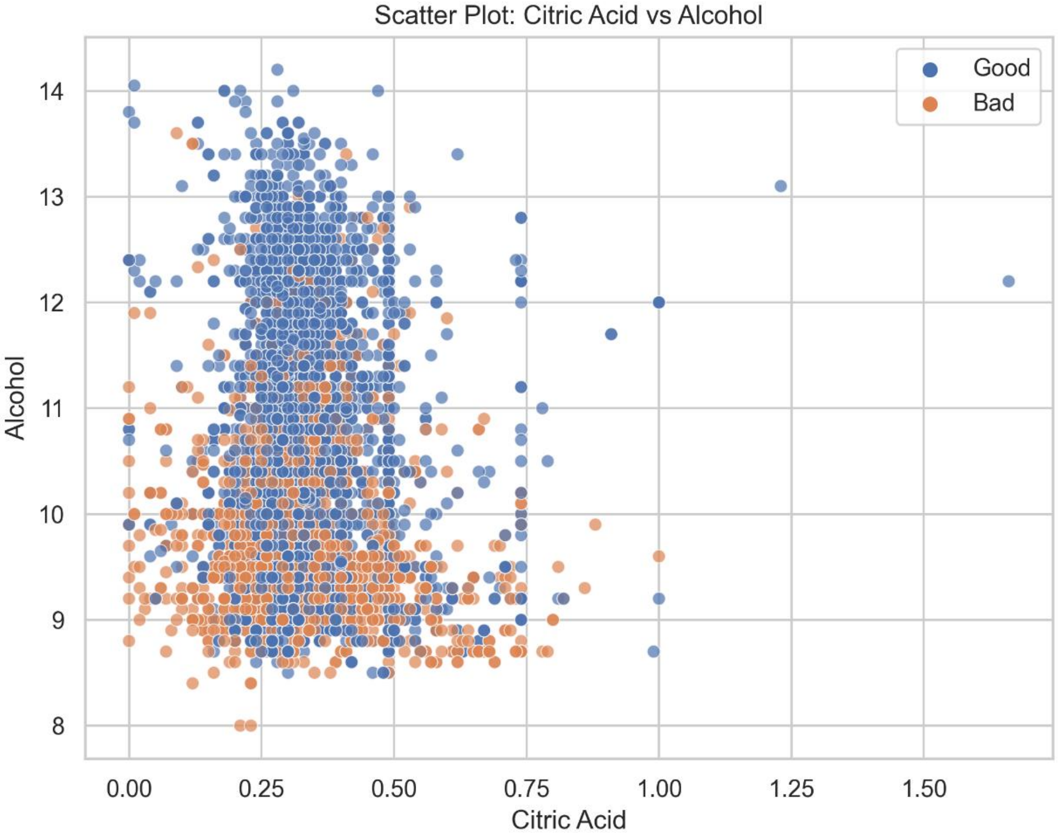
7. Scatter Plot: Volatile Acidity vs Alcohol:

Interpretation: This scatter plot displays the relationship between volatile acidity and alcohol content. The points exhibit a scattered pattern, suggesting a weak correlation. Wines with varying volatile acidity levels also have diverse alcohol content.



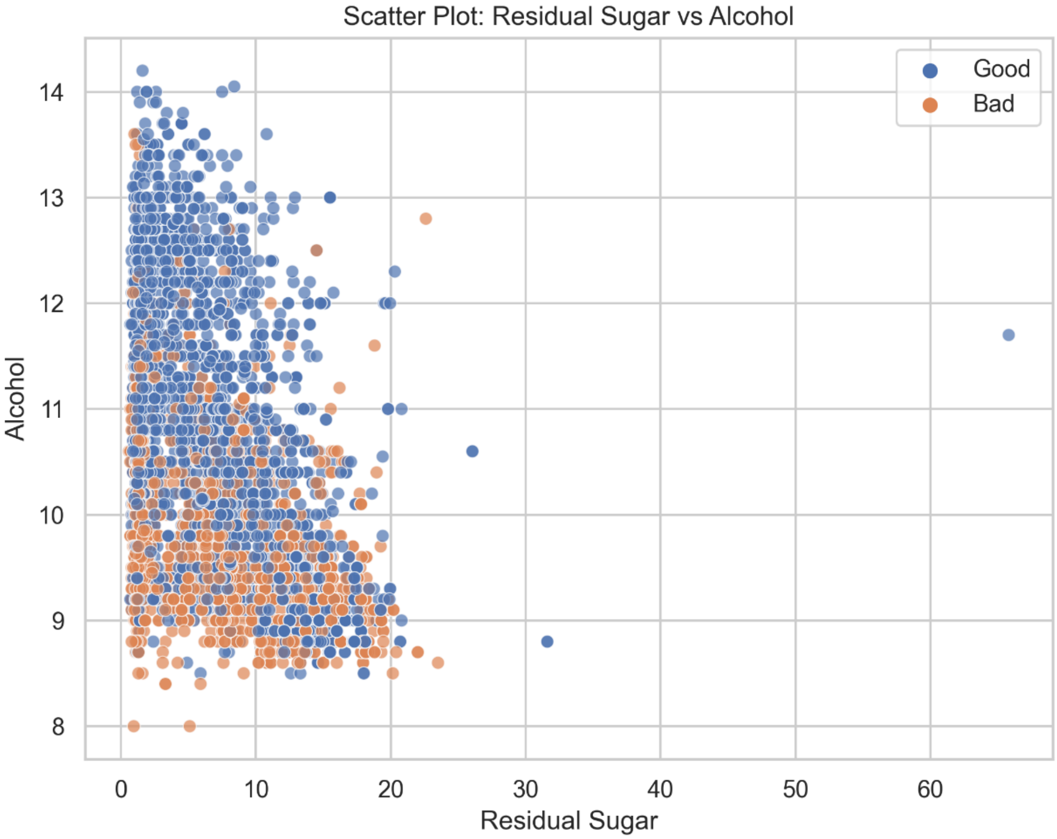
8. Scatter Plot: Citric Acid vs Residual Sugar:

Interpretation: The scatter plot showcases the relationship between citric acid and residual sugar. The points don't exhibit a strong pattern, indicating a weak correlation. Wines with varying citric acid levels also have diverse distributions of residual sugar.



9. Scatter Plot: Citric Acid vs Alcohol:

Interpretation: This scatter plot shows the relationship between citric acid and alcohol content. The points don't display a clear correlation. Wines with different citric acid levels also have diverse alcohol content.

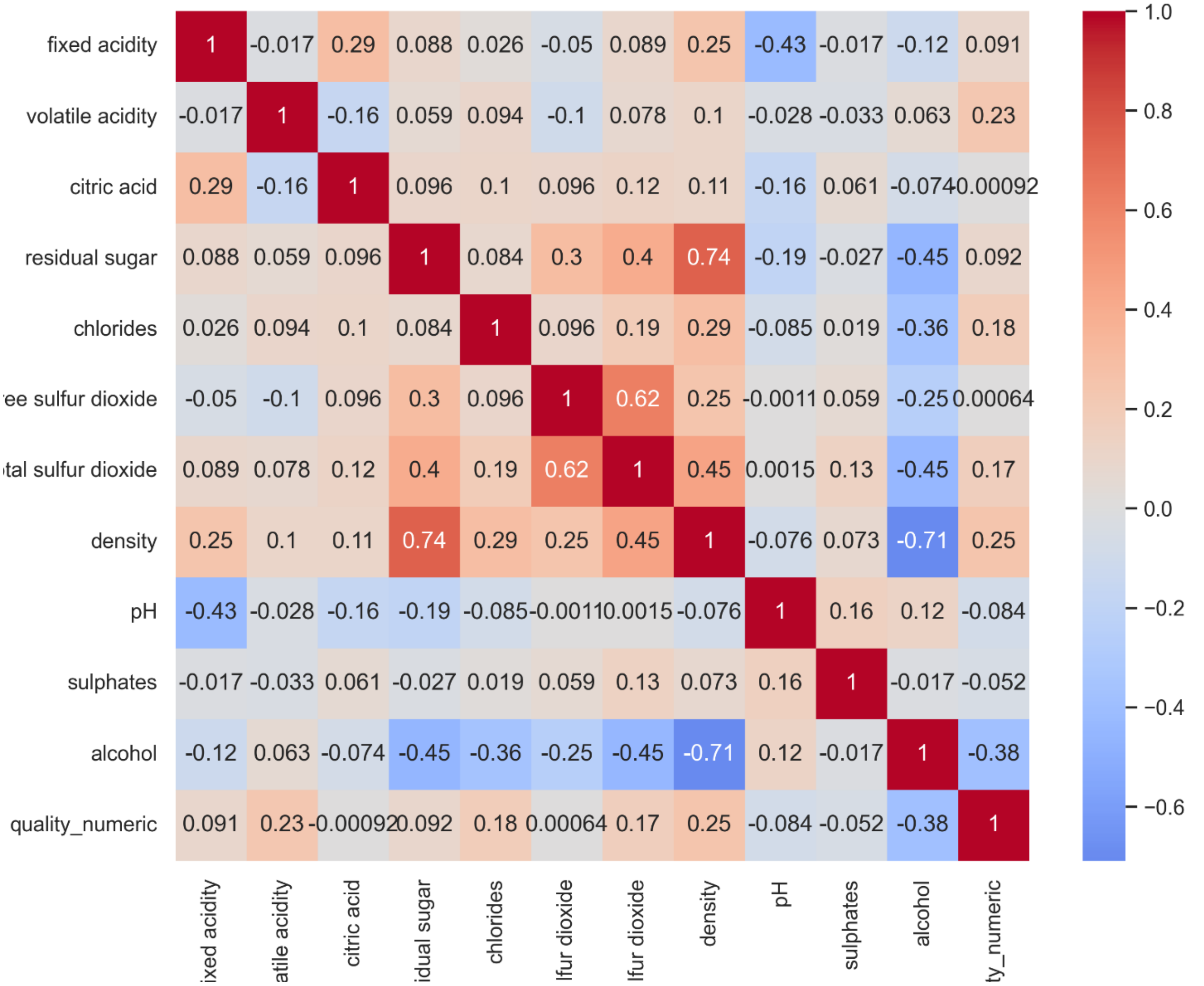


10. Scatter Plot: Residual Sugar vs Alcohol:

Interpretation: The scatter plot illustrates the relationship between residual sugar and alcohol content. The points show a scattered pattern, suggesting a weak correlation. Wines with varying residual sugar levels also have diverse alcohol content.

Conducting EDA for the Selected Study Data

Heatmap



Interpretation of the Correlation Heatmap:

The correlation heatmap provides insights into the relationships between numeric variables in the adjusted white wine quality dataset. The color scale ranging from cool to warm indicates the strength and direction of correlation.

1. **Fixed Acidity and Other Variables:**
    - Fixed acidity shows a moderate positive correlation with citric acid and density.

- There is a weak negative correlation with volatile acidity.

2. **Volatile Acidity and Other Variables:**
    - Volatile acidity exhibits a weak positive correlation with total sulfur dioxide and density.
    - It has a weak negative correlation with citric acid and alcohol.
3. **Citric Acid and Other Variables:**
    - Citric acid has a moderate positive correlation with fixed acidity and density.
    - It exhibits a weak positive correlation with pH.
4. **Residual Sugar and Other Variables:**
    - Residual sugar doesn't show strong correlations with other variables in the dataset.
5. **Chlorides and Other Variables:**
    - Chlorides exhibit a weak positive correlation with density and total sulfur dioxide.
6. **Free Sulfur Dioxide and Other Variables:**
    - Free sulfur dioxide doesn't show strong correlations with other variables.
7. **Total Sulfur Dioxide and Other Variables:**
    - Total sulfur dioxide shows a weak positive correlation with free sulfur dioxide and chlorides.
    - It has a moderate positive correlation with density.
8. **Density and Other Variables:**
    - Density displays a moderate positive correlation with fixed acidity, citric acid, and total sulfur dioxide.
    - It has a weak positive correlation with volatile acidity and chlorides.
9. **pH and Other Variables:**

- pH doesn't show strong correlations with other variables.

10. **Sulphates and Other Variables:**

    - Sulphates exhibit a weak positive correlation with alcohol.

11. **Alcohol and Other Variables:**

    - Alcohol has a weak positive correlation with sulphates.
    - It has a moderate negative correlation with density.

The correlation heatmap aids in identifying potential multicollinearity between variables, helping to determine which features might influence wine quality. However, correlation doesn't imply causation, so further analysis is needed to make meaningful conclusions about the relationships between variables.

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to isolate the individual effects of these variables on the dependent variable. From the correlation heatmap of the adjusted white wine quality dataset, we can identify potential multicollinearity by looking for pairs of variables with high correlation coefficients (close to 1 or -1). Here are some variables that might exhibit multicollinearity based on the heatmap:

1. **Density and Fixed Acidity:**

    - Density and fixed acidity have a moderate positive correlation.
    - This could suggest that wines with higher fixed acidity tend to have slightly higher density.

2. **Density and Citric Acid:**

    - Density and citric acid also have a moderate positive correlation.
    - This could indicate that wines with higher citric acid content tend to have slightly higher density.

3. **Total Sulfur Dioxide and Free Sulfur Dioxide:**

    - Total sulfur dioxide and free sulfur dioxide are moderately positively correlated.

- This could imply that wines with higher levels of total sulfur dioxide also have higher levels of free sulfur dioxide.

Code of the Heatmap

```
## Continuation
# Remove the 'quality' column temporarily
quality_column = df2['quality']
df2 = df2.drop(columns=['quality'])

# Correlation Heatmap
correlation_matrix = df2.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0)
plt.title('Correlation Heatmap')
plt.show()
# Add back the 'quality' column
df2['quality'] = quality_column
```

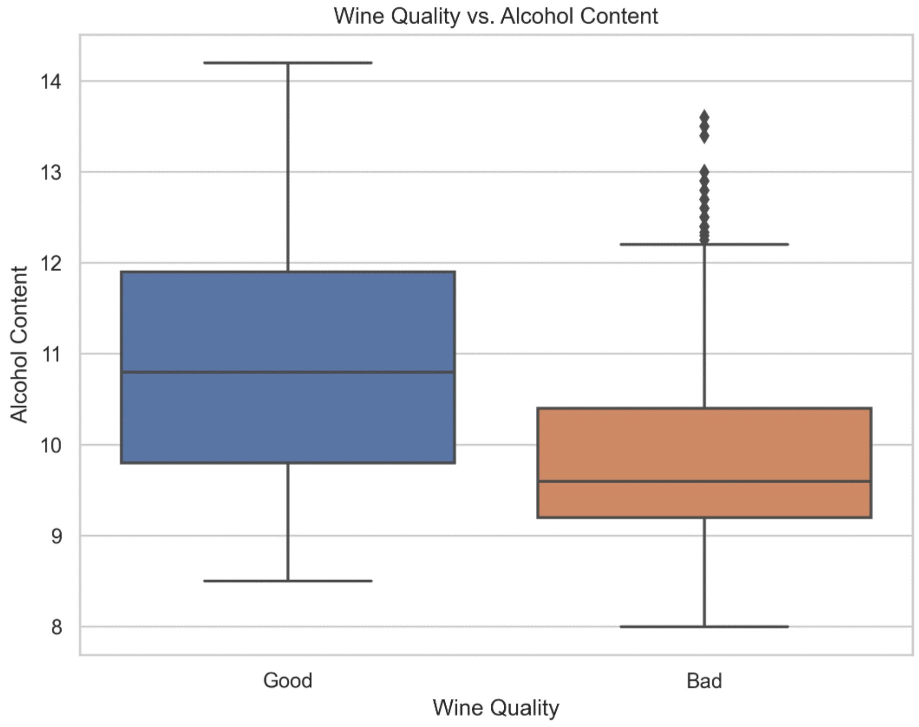
Box Plot for Wine Quality vs. Alcohol Content

Code:

```
## Continuation
# Box Plot for Wine Quality vs. Alcohol Content
plt.figure(figsize=(8, 6))
sns.boxplot(x='quality', y='alcohol', data=df2)
```

```
plt.title('Wine Quality vs. Alcohol Content')
plt.xlabel('Wine Quality')
plt.ylabel('Alcohol Content')
plt.show()
```

Output:



The box plot illustrates the relationship between wine quality and alcohol content in the white wine quality dataset.

- **Wine Quality Categories (X-Axis):** The x-axis represents the different categories of wine quality, which have been adjusted to 'Good' and 'Bad' based on the given criteria (Good >5, Bad<=5).
- **Alcohol Content (Y-Axis):** The y-axis represents the alcohol content of the wines.

Interpretation:

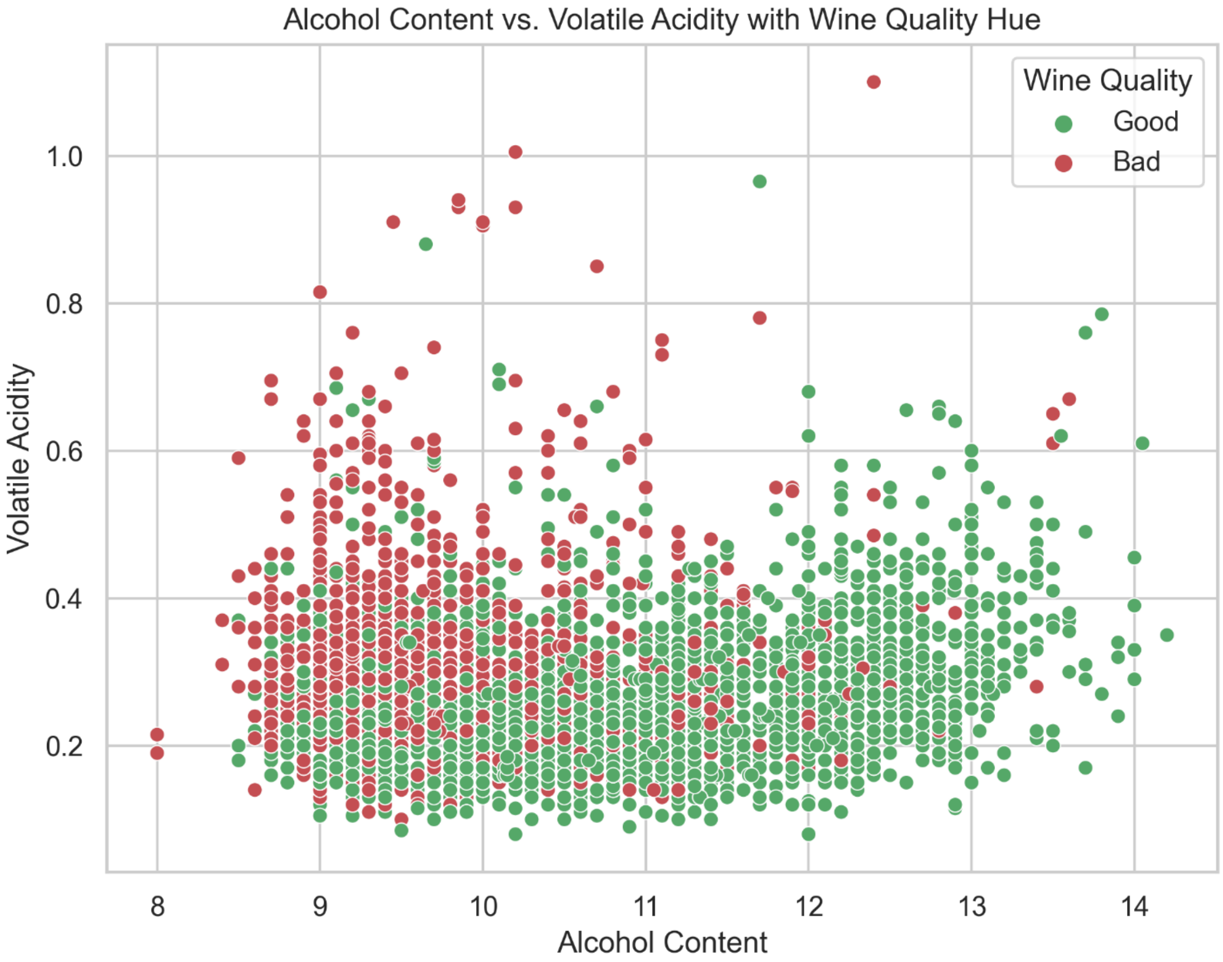
- The box plots show the distribution of alcohol content for wines categorized as 'Good' and 'Bad' quality.
- Wines with 'Good' quality tend to have a higher median alcohol content compared to wines with 'Bad' quality.
- The interquartile range (IQR) for 'Good' quality wines is larger, indicating a wider spread of alcohol content.
- 'Bad' quality wines generally have a lower median alcohol content and a narrower IQR.
- Outliers are present in both categories, suggesting some variability in alcohol content for both 'Good' and 'Bad' quality wines.

Scatter Plot for Alcohol Content vs. Volatile Acidity

Code:

```
## Continuation

# Scatter Plot for Alcohol Content vs. Volatile Acidity
plt.figure(figsize=(8, 6))
sns.scatterplot(x='alcohol', y='volatile acidity', data=df2, hue='quality', palette={'Good': 'g', 'Bad': 'r'})
plt.title('Alcohol Content vs. Volatile Acidity with Wine Quality Hue')
plt.xlabel('Alcohol Content')
plt.ylabel('Volatile Acidity')
plt.legend(title='Wine Quality')
plt.show()
```

Output:


Alcohol Content vs. Volatile Acidity with Wine Quality Hue
Wine Quality
Good
Bad
Volatile Acidity
1.0
0.8
0.6
0.4
0.2
8
9
10
11
12
13
14
Alcohol Content

The scatter plot visualizes the relationship between alcohol content and volatile acidity in the adjusted white wine quality dataset, color-coded by wine quality.

- Alcohol Content (X-Axis): The x-axis represents the alcohol content of the wines.
- Volatile Acidity (Y-Axis): The y-axis represents the volatile acidity of the wines.
- Color Hue (Legend): The legend indicates the wine quality categories ('Good' and 'Bad') through color differentiation.

Interpretation:

- The scatter plot allows us to observe the distribution of alcohol content and volatile acidity for wines of different qualities.
- Wines with 'Good' quality (green points) tend to have higher alcohol content and lower volatile acidity.
- 'Bad' quality wines (red points) generally exhibit lower alcohol content and higher volatile acidity.
- There is a trend showing a negative correlation between alcohol content and volatile acidity, especially for 'Bad' quality wines.
- Some 'Good' quality wines have lower alcohol content but still manage to maintain low volatile acidity.

Pair Plots

To decide which variables to include in the pair plots with different colors for wine qualities ('Good' and 'Bad') in the adjusted white wine quality dataset, I considered the variables that showed meaningful relationships and potential impact on wine quality. Based on my earlier analysis, I decided to include the following variables:

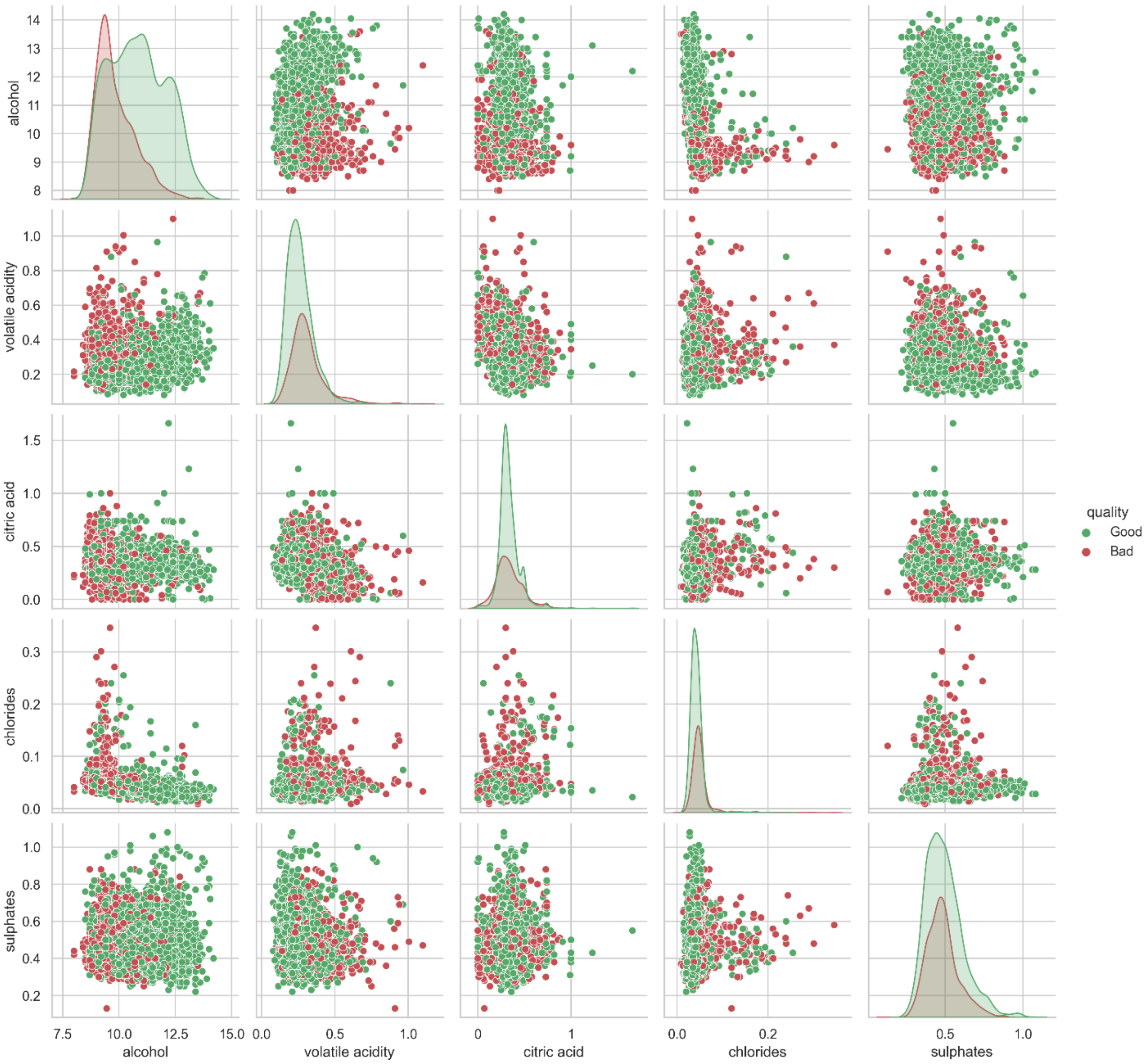
1. **Alcohol:** It showed a clear relationship with wine quality, with 'Good' quality wines having higher alcohol content.

2. **Volatile Acidity:** It exhibited differences between 'Good' and 'Bad' quality wines, making it an important variable.
3. **Citric Acid:** While there's some overlap, it still seemed to contribute to distinguishing between the quality categories.
4. **Chlorides:** It had slight visual trends suggesting its possible role in differentiating wine qualities.
5. **Sulphates:** Similar to chlorides, it might have some impact on wine quality differentiation.

Code:

```
## Continuation

# Selecting variables for pair plots
selected_columns = ['alcohol', 'volatile acidity', 'citric acid', 'chlorides', 'sulphates']


# Pair Plots with Hue for Wine Quality
sns.pairplot(df2[selected_columns + ['quality']], hue='quality', palette={'Good': 'g', 'Bad': 'r'})
plt.suptitle('Pair Plots with Hue for Selected Variables and Wine Quality', y=1.02)
plt.show()
```

Output:



Interpretation of Pair Plots with Hue for Selected Variables and Wine Quality:

The pair plots with hue provide a detailed visual exploration of the relationships between selected variables (Alcohol, Volatile Acidity, Citric Acid, Chlorides, and Sulphates) and their impact on wine quality ('Good' and 'Bad') in the adjusted white wine quality dataset.

1. **Diagonal Plots (Distribution Plots):**

    - Along the diagonal, distribution plots show the distribution of each selected variable.
    - For instance, the histogram for 'alcohol' reveals that most wines have alcohol content between approximately 8.5 and 12.5.

2. **Scatter Plots with Hue (Variable Relationships and Quality):**
    - The scatter plots illustrate the relationships between pairs of selected variables.
    - Points are color-coded based on wine quality, with 'Good' quality wines in green and 'Bad' quality wines in red.

Interpretation of Variable Relationships:

- **Alcohol Content vs. Other Variables:**
    - Wines with higher alcohol content tend to be associated with 'Good' quality wines.
    - The scatter plots show that 'Good' quality wines (green points) are concentrated in the upper range of alcohol content, while 'Bad' quality wines (red points) are more dispersed.
- **Volatile Acidity vs. Other Variables:**
    - 'Bad' quality wines appear to have higher volatile acidity compared to 'Good' quality wines.
    - Some overlap exists, indicating that volatile acidity alone might not solely determine wine quality.

- **Citric Acid, Chlorides, and Sulphates:**
    - Visual patterns show variations between 'Good' and 'Bad' quality wines.
    - However, the separation is less distinct than observed for alcohol content and volatile acidity.

Key Observations:

- The pair plots with hue provide a comprehensive view of how the selected variables interact with each other and their impact on wine quality.
- While some variables show clear separation between quality categories, others exhibit more complex relationships.
- The visualizations provide insights into how different combinations of variables might contribute to wine quality determination.

These pair plots offer valuable insights for understanding the relationships between the selected variables and wine quality categories, aiding in making informed decisions for further analysis and modeling.